Building pipeline-based NLP systems for your applications

Hua Xu

School of Biomedical Informatics,
University of Texas Health Science Center at Houston

Disclosure

- I receive grant funding from:
  - NIH: NLM, NIGMS, NCI
  - CPRIT (Cancer Prevention and Research Institute of Texas)
- I have been a consultant for:
  - Hebta LLC

What Is NLP?

- Broad Definition – any system that manipulates text or speech. It could involve various degrees of linguistic knowledge.
- NLP Systems
  - Natural Language understanding
  - Natural Language extraction
  - Natural Language generation
  - Machine translation
  - NLP-based information retrieval
  - NLP-interfaces

Study of Natural Language

- Human language (vs. formal and computer language)
- Linguistics - a description of language - used by theoretical linguists.
- Psycholinguistics - a cognitive model of how people understand and generate language.
- Computational linguistics - build computational models to understand and generate language.

Computational Linguistics

- An interdisciplinary field dealing with the statistical and/or rule-based modeling of natural language from a computational perspective
  - Driven by need to process natural language – convert to structured form for further computerized processes
  - Computational model is not necessarily same as human model - we don't understand much about human language facility

Overview of Linguistic Levels

- **Phonology**: units of sound combine to produce words (will not cover)
- **Morphology:** basic units combine to produce words
- **Lexicography**: syntactic (part of speech) and semantic categories of words
- **Syntax:** structures combine to produce sentences
- **Semantics:** meaning/interpretations
- **Discourse** – previous information affects the interpretation of the current information
- **Pragmatic:** context or world knowledge affects the interpretation of meaning

Morphology

- Definition: The study of how words are composed from smaller, meaning-bearing units (morphemes)
  - Inflection: Word stem + grammatical morpheme
    - like → likes, liked, liking
  - Derivation: Word stem + syntactic/grammatical morpheme
    - generalize → generalization
  - Compounding: Two base forms join to form a new word
    - bedtime
- Application: spelling check, stemming, POS tagging, speech recognition

Lexicography - Words

- Recognize word – Tokenization (determine the word boundary)
- Identify word – Lookup (map to dictionary entry)
- Categorize word – Tagging
  - Syntactic – Assign Part-of-Speech Tags
  - Semantic – Assign semantic categories

Syntax - Sentences

- ♦ Definition: study of the structure of a sentence.
  - – Categories combine with others to produce a well-formed structure with underlying relations
- ♦ Difficulties: ambiguous, nesting, omitted structures
  - – pain in (hands and feet) vs. (pain in hands) and fever
- ♦Parsing – determining syntax
  - – Formalisms: regular expressions vs. context-free grammar
  - – Partial vs. full parsing

Semantics

- ♦ Lexical level – to determine the meaning of a word
  - ♦ Semantic categories of a word
    - *Abdomen* – body location
    - *Fever* – symptom
    - *pt* – labtest (prothrombin time assay) vs. treatment (physical therapy)
  - ♦ Word sense disambiguation
- ♦ Grammatical level - word senses in a structure combine to form a meaning of the whole structure

Discourse

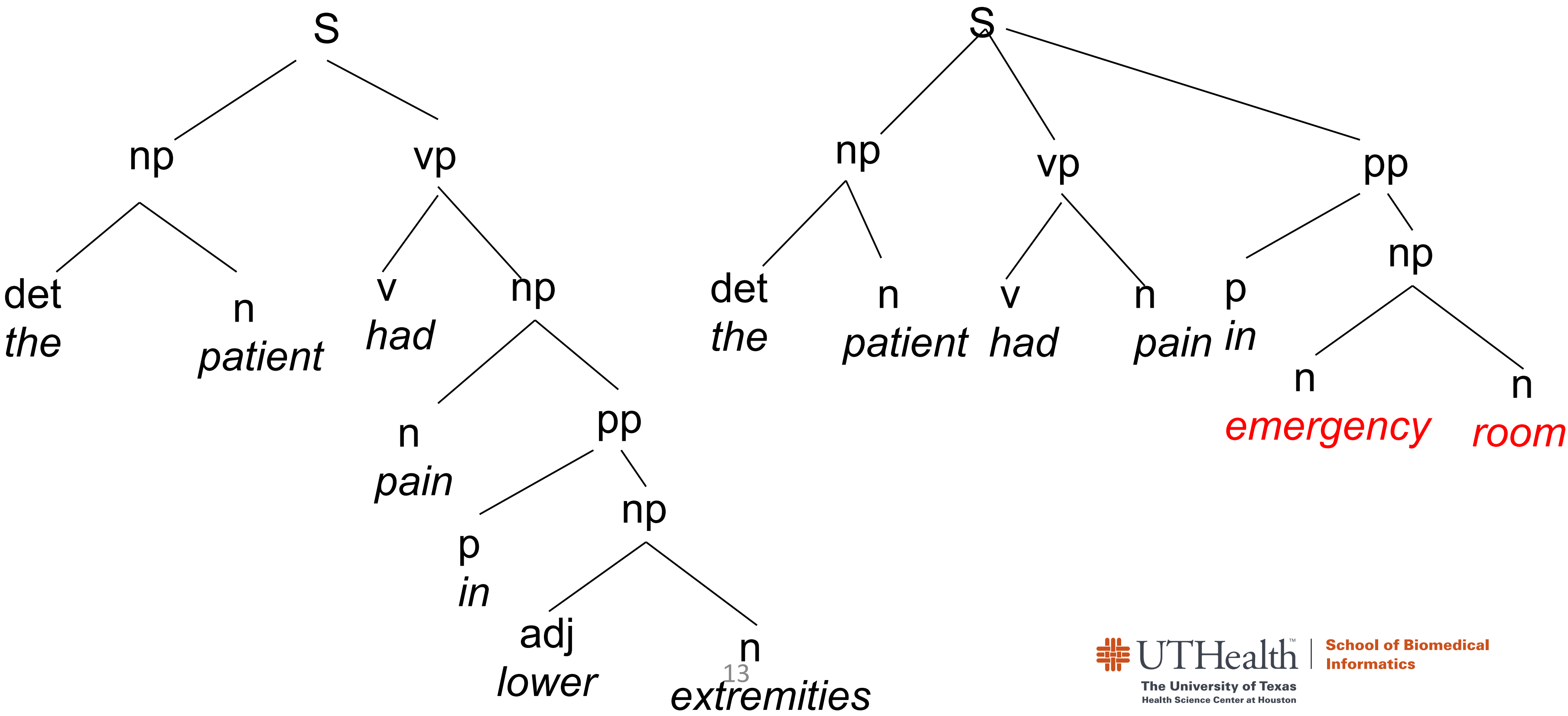
- ♦ Previous information in text affects current text
  - – Correct reference for pronouns, definite noun phrases, bridging noun phrases.
    - *Mass noted in left upper lobe. It was well-marginated.*
  - – Time of events
  - – Determining topic
  - – Coherence of text

Pragmatics

- Context affect meaning
  - Domain: *A mass was observed*
  - Section of Report: past history vs. hospital course
  - Prior information
- World knowledge affects interpretation
  - *He couldn't do any trading on the past Monday. (Market was closed on President Day - Monday.)*

It’s all about Ambiguity!

- POS tagging - saw (noun vs. verb)
- Semantic tagging - pt (patient, physical therapy, prothrombin time assay)
- Syntactic parsing - *The patient had pain in lower extremities.* *vs.*
  The patient had pain in emergency room.



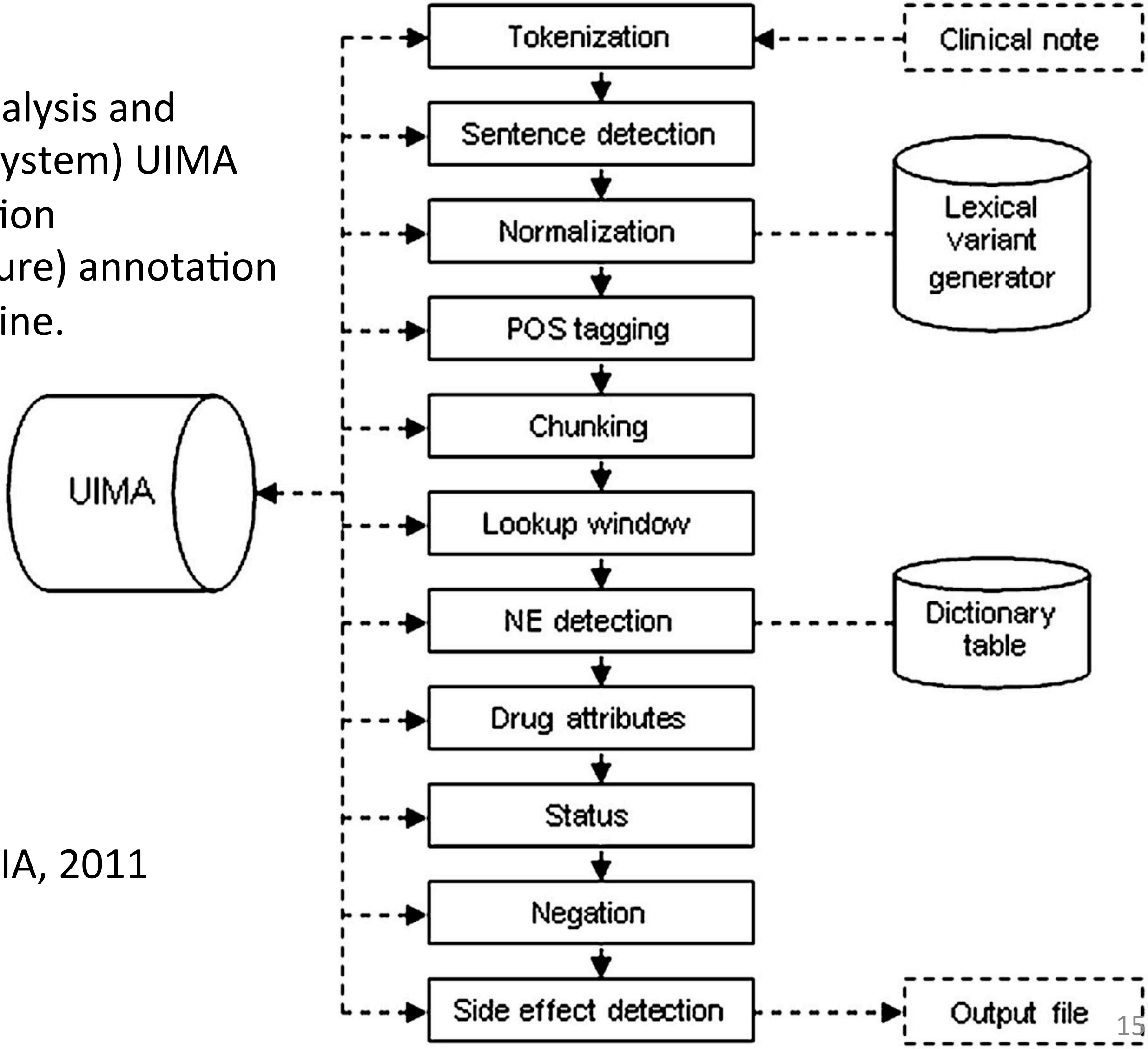
UTHealth | School of Biomedical Informatics
The University of Texas
Health Science Center at Houston

Most of current clinical NLP systems are information extraction systems

- General-purpose
  - MedLEE
  - MetaMap
  - cTAKES
  - KnowledgeMap Concept Identifier
  - ….
- Specific-purpose
  - MIST – the MITRE identification scrubber toolkit
  - MedEx – medication information extraction
  - ……

Pipeline-based architecture

cTAKES (clinical Text Analysis and Knowledge Extraction System) UIMA (Unstructured Information Management Architecture) annotation flow of side effect pipeline.



Source: Sohn et al. JAMIA, 2011

Demo of building clinical NLP pipelines using CLAMP

- Clinical Language Annotation, Modeling, and Processing Toolkit (CLAMP)
- Demo 1 – determine smoking status using rule-based approaches
- Demo 2 – extract lab names using a hybrid approach that combines machine learning and rules

Introduction to CLAMP

- A general purpose clinical NLP system built on proven methods

NLP Tasks		Ranking
Named entity recognition	2009 i2b2, medication	#2
	2010 i2b2 problem, treatment, test	#2
	2013 SHARe/CLEF abbreviation	#1
UMLS encoding	2014 SemEval, disorder	#1
Relation extraction	2012 i2b2 Temporal	#1
	2015 SemEval Disease-modifier	#1
	2015 BioCREATIVE Chemical-induced disease	#1

- An IDE (integrated development environment) for building customized clinical NLP pipelines via GUIs
  - Annotating/analyzing clinical text
  - Training of ML-based modules
  - Specifying rule

What does CLAMP address?

- The Transportability Problem of NLP
  - From one type of clinical notes to another
  - From one institute to another
  - From one application to another
- Need a solution for non-NLP experts to efficiently build high-performance NLP modules for individual applications!

CLAMP Demo 1

- Build a rule-based system to extract smoking status from clinical text
- Input: sentences containing patient smoking information
- Output: three types of status for each smoking mention:
  - Current Smoker: She has a prior history of smoking although not currently
  - Past Smoker: She is continuing to smoke
  - Non-Smoker: She denies any tobacco use , alcohol use

CLAMP Demo 2

- Build a hybrid (machine learning + rules) system for extracting lab test concepts from clinical text
- Input: discharge summaries
- Output: lab test concepts mentioned in the text with attributes of:
  - Offsets
  - Negation
  - UMLS CUIs

CLAMP Availability

- CLAMP is available in two versions:
  - CLAMP CMD (free)
  - CLAMP GUI (depends on the license)

  https://sbmi.uth.edu/ccb/resources/clamp.htm
- It is not an open source software, but source codes are available for collaborators with appropriate licenses.
- We are looking for collaborators to co-develop the system! If interested, please contact: Hua.Xu@uth.tmc.edu

Thank you!

Questions?

hua.xu@uth.tmc.edu